Proceedings of Machine Learning Research 260, 2024 ACML 2024

Orchestrating Plasticity and Stability: A Continual Knowledge Graph Embedding Framework with Bio-Inspired Dual-Mask Mechanism

Ailin Song[1] SONGAILIN@GDIIST.CN
Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, China

Yuhong Chen[1] YHCHEN2320@163.COM
College of Computer and Data Science, Fuzhou University, Fuzhou, China

Yusong Wang WANGYI@LR.PI.TITECH.AC.JP
Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan

Shuai Zhong[*] ZHONGSHUAI@GDIIST.CN
Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, China

Mingkun Xu[*] XUMINGKUN@GDIIST.CN
Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Learning in biological systems involves the intricate modeling of diverse entities and their interrelations, leading to the evolution of logical knowledge networks with accumulating experience. Analogously, knowledge graphs serve as semantic representations of entity relationships, playing a vital role in natural language processing and graph representation learning. However, contemporary knowledge graph embedding models often neglect real-world event updates, while existing continual knowledge graph research predominantly relies on conventional learning methods that inadequately leverage graph structure, thereby compromising their continual learning capabilities. This study introduces a novel **C**ontinual **m**ask **K**nowledge **G**raph **E**mbedding framework (CMKGE), designed to address these limitations. CMKGE integrates semantic attributes, network structure, and continual learning mechanisms to capture the dynamic evolution of knowledge. Inspired by biological signal propagation and Dale's principle, we introduce a dual-mask mechanism for neuronal inhibition and activation. This mechanism automatically filters critical old knowledge, enhancing model plasticity and stability. Through comprehensive evaluations on four datasets, we demonstrate CMKGE's superiority over state-of-the-art continual embedding models.

Keywords: Knowledge graph, Knowledge graph embedding, Graph Continual learning

1. Introduction

For humans, learning is an abstract modeling of intricate entities and relationships Illeris (2018) organized by the graph structure. And the way of learning is always accumulated. Similarly, the knowledge graph(KG) is a semantic network recording the complex relation-

[1]The two authors contribute equally to this work.
[*]Corresponding authors.

ships between different entities in the real world Zou (2020). Some datasets like WordNet, a comprehensive English lexicon, it includes the intricate semantic relationships among diverse words Fellbaum (1998). Specifically, the KG dataset comprises triplets(head entity, relationship, tail entity), where the relationship is unidirectional, such as (Beijing, located in, China). Meanwhile, knowledge graph embedding is a reasonable technique to project nodes and relationships into high-dimensional vector spaces to encode the relationships between entities numerically Liu et al. (2023). The application of knowledge graph embedding is vast and diverse, including various fields such as natural language processing Wang et al. (2021, 2023b), information retrieval Zheng et al. (2020); Li et al. (2023), and recommender systems Mezni et al. (2021). Human has strong adaptability to continually update, accumulate, and exploit knowledge Wang et al. (2023a). Similarly, the knowledge graph which has the dynamism of entities and relationships in the real world is continually expanding. Facing the great information of knowledge graph, retraining has huge time and computation consumption when the data is dynamically updated. Therefore the research on continual learning of knowledge graph has great significance Daruna et al. (2021).

However, neural networks Neural is difficult to handle continual data streams because new data learning often affects old data, leading to a deterioration in the processing effectiveness of old data. Specifically, learning new data may overwrite the old information. This phenomenon is catastrophic forgetting De Lange et al. (2021). The continual knowledge graph embedding learning is designed to mitigate catastrophic forgetting, effectively acquire novel knowledge, and preserve existing knowledge. Although, currently, several continual knowledge graph learning methods Omeliyanenko et al. (2023) have achieved significant success, two unresolved limitations remain. Firstly, these methods often design specific mechanisms to extract crucial prior knowledge, but they always lack a way to quantify the continual importance of embeddings. Second, they overlook the potential of explicit graph structures in accurately representing the dynamic nature of the evolving knowledge graph, only using the local semantic information of triples.

To deal with these limitations, we propose a biologically plausible framework called **C**ontinual **M**asked **K**nowledge **G**raph **E**mbedding learning (CMKGE). This framework is mainly composed of four modules, bio-inspired masked data filtering, local semantic attributes, global graph network structure, and knowledge continual learning to orchestrate plasticity and stability. CMKGE has two well-designed components: the first component designed to preserve the stability by leveraging the randomness and controllability of synaptic excitation and inhibition, thereby automatically discriminating between crucial and less significant knowledge, which design is inspired by the finding of Ipsen and Peterson Ipsen and Peterson (2020) that synaptic excitation and inhibition plays a crucial role in the processing and regulation of information. The second component combines local semantic attributes and global semantic structures within the knowledge graph to represent explicit graph structures. The integration of local attributes and global structure is effective. It not only learning but also integrating previous and new knowledge in the knowledge continual learning framework, thereby facilitating a balance between plasticity and stability.

Specifically, from randomness and control, we set up a novelty dual-mask mechanism with learnable weights to facilitate the selective filtering of significant historical knowledge, based on spontaneous asynchronous irregularities in neural activity Roland (2017) and controlled states of excitation and inhibition of synapses Denève et al. (2017). This mechanism accommodates the integration of new important information, resulting in a robust and adaptive continual knowledge graph representation. Then, we capture and fuse both local and global semantic attributes as the node representation. In continual knowledge learning, the importance of previous representations is determined not only by neighbor information but also by their position with the structure of the knowledge graph. Furthermore, we use weighted regularization to punish knowledge change. This design enables our framework to capture the contextual relevance of representations, bolstering its ability to retain significant historical knowledge while accommodating new information.

Our main contributions are summarized as follows:

- CMKGE proposes a novelty framework. Learning continual knowledge embedding from four modules: data filtering, knowledge fusion, knowledge learning, and knowledge forgetting to orchestrate plasticity and stability.
- CMKGE innovatively realizes automatic parameter filtering from the biological perspective in continual knowledge graph learning to effectively enhance the robustness of the framework, achieving the filter of important embeddings.
- CMKGE combines the information of the whole graph structure with fine-grained local attributes, enhancing the feature representations to retain both historical and new knowledge.
- CMKGE outperforms state-of-the-art(SOTA) methods in link prediction on four benchmarks across a range of comprehensive experiments, demonstrating its effectiveness.

2. Related Work

2.1. Knowledge Graph Embedding

Knowledge graph embedding transforms entities and their relations into a vector space, representing head, tail entities, and relations as vectors that capture their inherent connections. At present, the mainstream methods of knowledge graph embedding learning include translation, bilinear, neural network, rotation and so on. Translation models, such as TransE Bordes et al. (2013), TransH Wang et al. (2014), etc. define relationships as translational shifts between head and tail entities. Bilinear models compute the confidence of the semantics of entities and relations in vector Spaces, including models such as RESCAL Nickel et al. (2011), DisMult Yang et al. (2014), ComplEx Trouillon et al. (2016), etc. Neural network models such as ConvE Dettmers et al. (2018), CapsE Nguyen et al. (2018), etc. Combine the idea of convolution with embedded learning of KG. Rotation models treat relationships as rotations between head and tail entities, including RotatE Sun et al. (2019), QuatE Zhang et al. (2019), etc. Additionally, some recent methods have explored the utilization of graph neural networks to enrich the comprehension of knowledge graph embeddings Molokwu and Kobti (2021); Wang et al. (2024). Liang et al. Liang (2023) introduced a self-supervised learning method that integrates both the comprehensive graph structure and semantic information, resulting in robust and effectively embeddings.

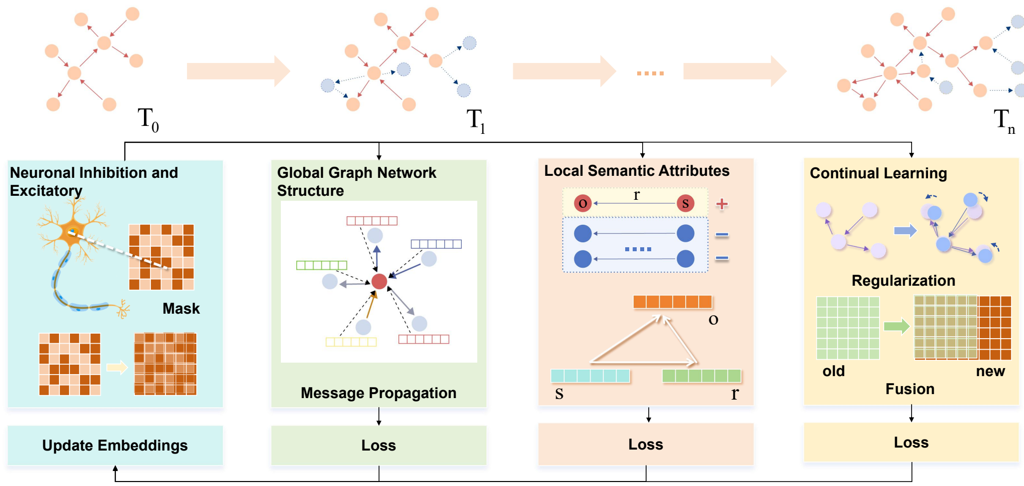

Figure 1: Schematic diagram of CMKGE framework. The framework consists of four parts. Neuronal Inhibition and Excitatory take advantage of the regulating effect of biological synapses on neuronal signal transmission. A learnable mask is designed to automatically extract key information of old knowledge and mask marginal information as the pre-input of new knowledge learning. Global Graph Network Structure and Local Semantic Attributes learn knowledge embedding from both global structure and local attributes. Continual Learning uses weighted regularization and knowledge fusion to mitigate the effects of catastrophic forgetting.

2.2. Continual Knowledge Graph Embedding

Continual learning requires constant adaptation to new information and tasks, thus eliminating the need to retrain from scratch. Recently, continual knowledge graph embedding methods can be categorized into three groups. Dynamic architecture approaches: They dynamically adapt their neural architecture to accommodate new information while retaining old parameters Zhang et al. (2023). By adjusting architecture, they respond to changes in the knowledge graph and incorporate new knowledge effectively. Memory approaches: These methods retain learned important knowledge, where past experiences are replayed during training to reinforce retention of old knowledge Liu et al. (2024); Omeliyanenko et al. (2023). Regularization approaches: These methods impose constraints on updating neural weights during training Yang et al. (2023); Chen et al. (2023). They expect the model to learn new information without forgetting important knowledge. Recently, Cui Cui et al. (2023) et al. proposed a lifelong knowledge graph embedding model called LKGE. It considers knowledge transfer and retention of the learning on growing snapshots of a KG without learning embeddings from scratch. It includes a masked KG autoencoder, an embedding transfer strategy and a regularization method. However, these methods are based on the conventional knowledge graph embedding techniques, which mainly rely on local triplet information while overlooking the crucial structural information of the entire KG. Furthermore, in the realm of traditional continual learning, these methods focus on capturing important old knowledge but overlook the significance of parameter selection.

3. Preliminaries

The knowledge graph dataset comprises triplets(head entity, relationship, tail entity), where the relationship is unidirectional. One prominent approach for knowledge graph embedding is TransE Bordes et al. (2013), which defines relations as translational transformations within the low-dimensional embeddings of entities. By employing vector addition, TransE effectively captures the interdependency among the three constituents of a triplet: the head entity, the relationship, and the tail entity, as demonstrated in Formula (1).

$$\mathbf{o} \approx \mathbf{s} + \mathbf{r}, \tag{1}$$

$\mathbf{s}$, $\mathbf{r}$, and $\mathbf{o}$ denote the vector of the head entity, relationship, and tail entity respectively.

In the context of traditional continual learning, regularization techniques play a significant role in enhancing the generalization ability of models and preventing overfitting. Regularization is typically achieved by imposing constraints on the model parameters or introducing additional penalty terms into the loss function. Two commonly employed regularization methods are $\mathcal{L}_1$ regularization and $\mathcal{L}_2$ regularization.

$$\mathcal{L}_1 = \sum_{i=1}^{N} |w_i| = \|\mathbf{W}\|_1, \tag{2}$$

$\mathcal{L}_1$ regularization as demonstrated in Formula (2) involves adding the absolute values of the weight parameters to the loss function. The model's weights follow a Laplace distribution, resulting in sparse parameters that are often beneficial for feature selection. Where w denotes the elements of weight vector $\mathbf{W}$, and N means the number of vector's elements.

$$\mathcal{L}_2 = \sqrt{\sum_{i=1}^{N} |w_i^2|} = \|\mathbf{W}\|_2, \tag{3}$$

On the other hand, $\mathcal{L}_2$ regularization as demonstrated in Formula (3) introduces the squared values of the weight parameters into the loss function. This approach encourages the model's weights to follow a Gaussian distribution, often leading to smaller parameter values. The resulting model is typically more robust and less prone to overfitting.

4. Methods

This method delves into knowledge embedding and accurately captures the dynamic evolution of knowledge through four interconnected dimensions: the fine-grained local semantic, the comprehensive global structure, the continual learning, and the biological neuron.

4.1. Notation

This section introduces the basic notation used in the method. Knowledge graph consists of individual triplets $\mathcal{T} = (s, r, o)$, where s denotes the head entity, r denotes the relationship, and o denotes the tail entity. As the continual learning process, the knowledge graph generates a sequence of knowledge snapshots $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, ...\mathcal{S}_t\}$. $\mathcal{S}_t$ represents the total dataset at moment t. We utilize graphs to describe the structure of the knowledge graph as $\mathcal{G}_t = \{\mathcal{V}_t, \mathcal{E}_t, \mathcal{X}_t, \mathcal{R}_t\}$. t represents the time scale, $\mathcal{V}$ means the set of nodes, and $\mathcal{E}$ means the set of edges. $\mathcal{X} \in \mathbb{R}^{n\times d}$ is the feature matrix of nodes, where n represents the number of vertex and d represents the embedding size. And $\mathcal{R} \in \mathbb{R}^{r\times d}$ denotes the feature matrix of edges where r represents the number of relations and d represents the embedding size.

4.2. Local Semantic Attributes

Firstly, to learn the fine-grained representations we refine our attention to a single snapshot, specifically on an individual triplet relationship. Inspired by the idea of the TransE Bordes et al. (2013), we envision this relationship as a translation pipeline connecting the head entity to the tail entity. The knowledge of multi-dimensional vectors of representation follows the principle of vector addition, as demonstrated in Formula (4).

$$\mathbf{H}_v = f(s, r) = \mathbf{H}_s + \mathbf{H}_r, \tag{4}$$

where $\mathbf{H}_s \in \mathbb{R}^{n\times d}$ is a vertex feature matrix, with each row corresponding to the features of the head node of the triples. $\mathbf{H}_r \in \mathbb{R}^{r\times d}$ as same as $\mathbf{H}_s$ is the features of the relation. $\mathbf{H}_v \in \mathbb{R}^{n\times d}$ is the feature matrix of node set $\mathcal{V}$ after calculating. Vertically concatenating triples (s, r, o) and their inverse relationships (o, r, s) in a dataset. n represents the node in the first column. r represents the edge of the second column. Therefore, their numbers are equal and satisfy the conditions for matrix addition. Then we construct a score function using $\mathcal{L}_1$ norm to compute the difference between the translated embeddings and tail entity embeddings, as demonstrated in Formula (5).

$$S(s, r, o) = \|\mathbf{H}_v - \mathbf{H}_o\|_1, \tag{5}$$

Furthermore, to enhance the robustness of the learned representations, we construct negative groups labeled for comparison. By evaluating the score difference between positive and negative groups, we compute the loss, as demonstrated in Formula (6).

$$\mathcal{L}_{local} = \frac{1}{|\mathcal{T}_t|} \sum_{(s,r,o)\in\mathcal{T}} max(0, S^- - S^+ + \alpha), \tag{6}$$

where α is the margin parameter. $|\mathcal{T}_t|$ denotes the number of triples at time t. The S^+ denotes the scores assigned to positive samples, whereas S^- corresponds to the scores of negative samples. Taking into account the aforementioned principles, the optimal scenario would involve minimizing S^+ towards zero while maximizing S^-, ensuring a clear separation between positive and negative representations.

4.3. Global Graph Network Structure

Furthermore, to facilitate a more comprehensive study of embeddings and mitigate the impact of negative triple groups, we adopt a global perspective in our learning approach. Considering the adjacency relations among entities within the overarching graph structure of the knowledge graph. We draw inspiration from Graph Convolutional Networks(GCN) and construct a message propagation network designed to aggregate information from neighboring entities for each entity, as elucidated in Formula (7).

$$\mathbf{H}_{fv} = f_{message}(\mathbf{A}, \mathbf{H}_v) = \sigma(\mathbf{D}^{-\mathbf{1}}\mathbf{A}\mathbf{H}_v), \tag{7}$$

where $\mathbf{H}_{fv} \in \mathbb{R}^{|\mathcal{T}_t| \times d}$ means the fused vertex features matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the adjacency matrices on one snapshot of knowledge graph. $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix recording the node degrees. $\sigma(\cdot)$ is the activate function.

Because there is no edge-relationship aggregation in GCN. The fusion of edge features inherit the idea of local semantic learning, as demonstrated in Formula (8).

$$\mathbf{H}_{fe} = f(s, o) = \mathbf{H}_s - \mathbf{H}_o, \tag{8}$$

where $\mathbf{H}_{fe} \in \mathbb{R}^{|\mathcal{T}_t| \times d}$ denotes the fused edge features matrix.

Considering the continual learning process, we not only account for newly emerging graph relationships but also seamlessly integrate them with the preexisting network for message aggregation, as detailed in Formula (9).

$$\mathbf{H}_t = \sigma(\mathbf{W} \cdot \mathbf{H}_{t-1} + (1 - \mathbf{W})\mathbf{H}_t). \tag{9}$$

$\mathbf{H}_t$ denotes the node feature matrix or edge feature matrix at the current moment in time. $\mathbf{H}_{t-1}$ denotes the previous moment. To mitigate the potential for catastrophic forgetting of prior knowledge by incoming knowledge embeddings, we adopt a weighted fusion vector $\mathbf{W} \in \mathbb{R}^{1 \times |\mathcal{T}_t|} = \{\mathbf{w}_1, \mathbf{w}_2, ...\mathbf{w}_i\}$, merging the embeddings of new knowledge with the previously learned old knowledge, as detailed in Formula (10).

$$\mathbf{w}_i = \frac{(\mathbf{d}_i)_{t-1}}{(\mathbf{d}_i)_{t-1} + (\mathbf{d}_i)_t}, \tag{10}$$

where $\mathbf{d}_i$ denotes the degree of vertex i. In this formula, we utilize the ratio of the vertex's degree at the previous time step to its current degree as the weight associated with vertex i. For edges, the weighting scheme remains analogous, with the weight reflecting the number of edges present. The define of loss as detailed Formula (11).

$$\mathcal{L}_{global} = \frac{\|\mathbf{H}_{fv} - \mathbf{H}_{\mathbf{v}}\|_2^2}{|\mathcal{V}_t|} + \frac{\|\mathbf{H}_{fe} - \mathbf{H}_e\|_2^2}{|\mathcal{E}_t|}, \tag{11}$$

where $|\mathcal{V}_t|$ and $|\mathcal{E}_t|$ represents the number of vertex and edge.

4.4. Continual Learning

In each individual learning snapshot, we focus on capturing both local and global perspectives of new knowledge. In the ever-expanding landscape of the knowledge graph, we

Datasets	$\mathcal{S}_1$			$\mathcal{S}_2$			$\mathcal{S}_3$			$\mathcal{S}_4$			$\mathcal{S}_5$		
	$\mathcal{T}_1$	$\mathcal{E}_1$	$\mathcal{R}_1$	$\mathcal{T}_2$	$\mathcal{E}_2$	$\mathcal{R}_2$	$\mathcal{T}_3$	$\mathcal{E}_3$	$\mathcal{R}_3$	$\mathcal{T}_4$	$\mathcal{E}_4$	$\mathcal{R}_4$	$\mathcal{T}_5$	$\mathcal{E}_5$	$\mathcal{R}_5$
ENTITY	46,388	2.909	233	72,111	5,817	236	73,785	8,275	236	70,506	11,633	237	47,326	14,541	237
RELATION	98,819	11,560	48	95,535	13,343	96	66,136	13,754	143	30,032	14,387	190	21,594	14,541	237
FACT	62,024	10,513	237	62,023	12,779	237	62,023	13,586	237	62,023	13,894	237	62,023	14,541	237
HYBRID	57,561	8,628	86	20,837	10,040	102	88,017	12,779	151	103,339	14,393	209	40,326	14,541	237

Table 1: A brief description of the datasets.

meticulously examine the intricate interplay between old and new knowledge. To mitigate the risk of catastrophic forgetting, we introduce a regularization mechanism imposing penalties on changes made by new knowledge to old knowledge, thereby preserving the integrity of previously learned concepts, as captured in Formula (12).

$$\mathcal{L}_{regulation} = \sum_{i=0}^{v} \|(w_v)_i(x_t - x_{t-1})\|_2^2 + \sum_{i=0}^{e} \|(w_r)_i(r_t - r_{t-1})\|_2^2. \tag{12}$$

x denotes the features vector of vertex, and r denodes the features of edge. We implement a weighted penalty to the alteration of the embeddings, thereby achieving a more harmonious balance between the retention of old knowledge and the assimilation of new knowledge. Without losing newly learned knowledge or adding to catastrophic forgettin. The define of $(w_v)_i$ and $(w_r)_i$ as same as the Formula (10).

4.5. Neuronal Inhibition and Excitatory

Undeniably, not all old weights are equally significant, identifying and filtering crucial parameters for propagation is imperative. Inspired by neuroscience, we introduce synaptic excitation and inhibition mask mechanisms to mimic the inherent randomness and plasticity of human cognition, as captured in Formula (13).

$$\mathbf{X}_{t+1} = \mathbf{M}_{mask}(\mathbf{X}_{old}\|\mathbf{X}_{new}). \tag{13}$$

$\mathbf{X}_{old} \in \mathbb{R}^{n_t \times d}$ is embeddedness of knowledge learned at the present moment, and $\mathbf{X}_{new} \in \mathbb{R}^{(n_{t+1}-n_t)\times d}$ is embeddedness of new knowledge entering at the next moment.

$$\mathbf{M}_{mask} = \mathbf{M}_{random} \cdot f_{heaviside}(\mathbf{X}_t - \beta). \tag{14}$$

The neuronal activity exhibits spontaneous, asynchronous, and irregular firing patterns Roland (2017), yet these are continually balanced through intricate excitatory and inhibitory mechanisms between neurons Denève et al. (2017). Drawing inspiration from biological randomness and controllability, we introduce a dual-mask mechanism that integrates random masking $\mathbf{M}_{random}$ and weighted masking function $f_{heaviside}(\mathbf{X}_t - \beta)$ strategies to mimic the knowledge learning functions of the human brain, ultimately aiming to mitigate the adverse effects of catastrophic forgetting. $\mathbf{M}_{random}$ is initialized by a full 1 matrix, with partial values assigned to 0 in a 10% ratio. $f_{heaviside}()$ is the Heaviside Step Function. $\mathbf{X}_t$ is the learned feature vector at the current moment. And β is a learnable threshold parameter.

These mechanisms make CMKGE to selectively emphasize and suppress weights. Not only augments the flexibility and precision of the learning process but also brings the knowledge-embedding procedure into closer alignment with the brain learning mechanisms.

Model	ENTITY				RELATION				FACT				HYBRID			
Metric	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
Fine-tune	0.168	0.087	0.190	0.323	0.089	0.039	0.101	0.183	0.178	0.095	0.199	0.347	0.133	0.069	0.147	0.256
PNN	0.229	0.131	0.264	0.423	0.168	0.098	0.192	0.306	0.160	0.088	0.189	0.292	0.186	0.104	0.215	0.348
CWR	0.089	0.028	0.115	0.205	0.022	0.010	0.024	0.043	0.086	0.032	0.098	0.194	0.040	0.016	0.050	0.081
SI	0.155	0.074	0.179	0.310	0.113	0.056	0.130	0.223	0.177	0.094	0.199	0.346	0.108	0.046	0.123	0.224
EWC	0.230	0.131	0.264	0.424	0.154	0.084	0.178	0.293	0.201	0.113	0.229	0.382	0.170	0.087	0.198	0.333
GEM	0.165	0.086	0.188	0.318	0.089	0.039	0.101	0.184	0.173	0.092	0.194	0.341	0.127	0.065	0.141	0.246
EMR	0.173	0.092	0.195	0.332	0.111	0.052	0.125	0.225	0.169	0.089	0.189	0.334	0.137	0.074	0.151	0.259
LKGE	0.234	0.136	0.269	0.425	0.180	0.096	0.207	0.347	0.209	0.121	0.235	0.386	0.201	0.114	0.230	0.374
CMKGE	**0.247**	**0.146**	**0.285**	**0.444**	**0.210**	**0.112**	**0.239**	**0.389**	**0.212**	**0.120**	**0.235**	**0.390**	**0.210**	**0.122**	**0.245**	**0.396**

Table 2: The performance of the tested four datasets.

5. Experiments Setting

5.1. Dataset

The datasets is FB15K-237 a subset extracted from the Freebase Knowledge Base, which is widely used in knowledge graphs. The datasets are divided into four sections Cui et al. (2023): ENTITY, RELATION, FACT, and HYBRID focusing on their respective dynamic growth. For each temporal snapshot, the training dataset, validation set, and test set are partitioned in a 3:1:1. **ENTITY** encapsulates the dynamic transformations of entity objects in the real world. **RELATION** encapsulates the dynamic semantic associations. **FACT** captures the dynamic evolution of knowledge triples. **HYBRID** is the randomized and dynamic variations in knowledge triples.

5.2. Comparison Algorithm

The details of the comparison algorithm are as follows, uniformly based on the TransE Bordes et al. (2013) model. **Fine-tuning.** Conversely, this strategy aims to maintain the integrity of all existing data. It selectively initializes only the parameters pertinent to the new triplet relationships. **Regularization.** We compared it with EWC and SI models. Both use loss functions to limit the updating of important parameters when learning new tasks. EWC directly penalizes all changes in old weights, while SI takes into account the importance of different parameters and minimizes changes in important parameters. **Replay.** GEM and EMR selectively store part of the data for use. In the GEM, this data is used to limit the gradient update of new tasks to ensure that the loss of old tasks does not increase. In EMR, this data is stored in a playback buffer and used for playback training when learning new tasks. **Dynamic Structure.** We compare with PNN and CWR methods. Both modify the model parameters and reduce the influence of new knowledge on the model parameters by selectively freezing the model parameters. Or to preserve the information of old parameters through the fusion of new and old model parameters to cope with catastrophic forgetting. **Continual Knowledge Graph Learning.** We compare it with the state-of-the-art continual Knowledge Graph Learning(LKGE) Cui et al. (2023) approach. This model includes a masked KG autoencoder for embedding learning and updating, with an embedding transfer strategy and an embedding regularization method.

5.3. Evaluation Metrics

We use MRR and Hit@n ($n \in \{1, 3, 10\}$) as the evaluation metrics. They are commonly used in recommender systems and link prediction. Specifically, MRR indicates Mean Reciprocal Rank. Hit@n($n \in \{1, 3, 10\}$) means the average percentage of triples that rank less than n in the link prediction, considering the n triples that are most related to entity1. Calculate the correct proportion. For fair comparison, we set the same batch size 2048, mask ratio of 0.001. Use Adam as an optimizer. The other hyperparameters are fine-tuned for best results. Embedding dimension in {100, 200}, learning rate in {0.001,0.0001}, and the proportion of regularization in the loss calculation in {0.1, 0.01}.

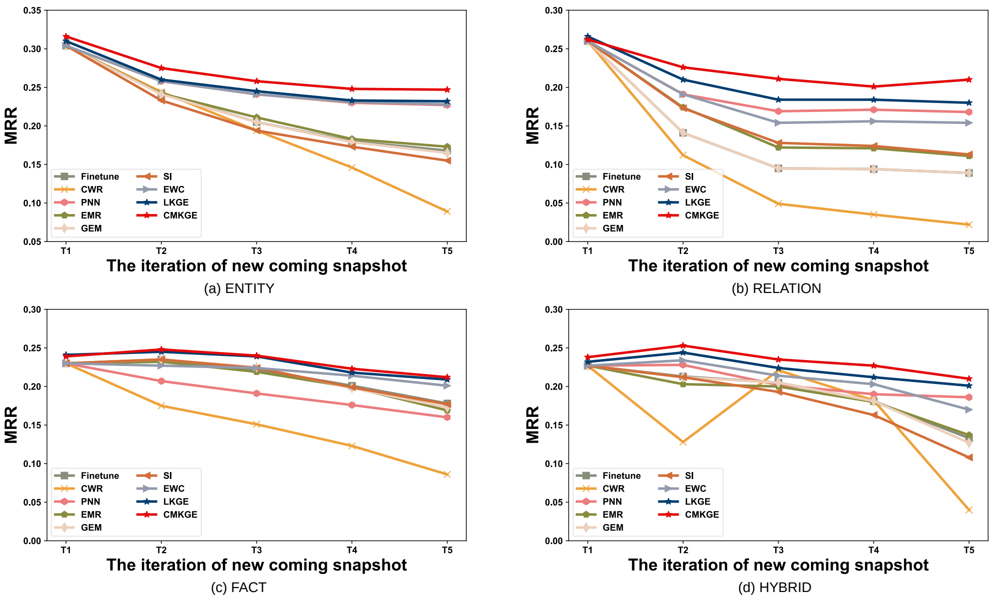

Figure 2: Continual learning ability of CMKGE on four datasets. CMKGE, the red line, achieves the best results in terms of smoothing and downward trend.

6. Experiments

The experiment mainly compares from four aspects: performance, continual learning ability, robustness test, and loss convergence.

6.1. Performance Comparison

We evaluate the effectiveness of ours compared with several classical and state-of-the-art methods in Table 2, where the best performance is highlighted in bold and the second-best results are underlined. From Table 2, the proposed method reflects a promising performance.

We introduce weighted regularization, which provides flexibility compared to EWC's direct penalty on weight changes. Compared to the biological mechanism of SI, the ability to enhance important weights. Our dual-mask mechanism selects important weights and

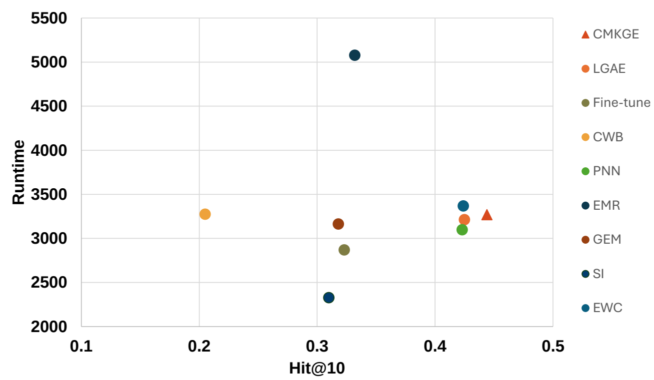

Figure 3: Performance and Consumption. The red triangle represents the coordinate position of our framework. Our framework is in the bottom right corner.

enhances their impact from the perspectives of randomness and controllability. It is more lightweight and efficient. Compared to replay-based GEM and EMR, we automatically filter out important information to retain. This mechanism employs updatable parameter control to enable the model to adaptively adjust the masking ratio, thus preserving crucial information, enhancing computational efficiency, and balancing the plasticity and stability. Compared with the dynamic structure, instead of freezing the old weights, we effectively fuse the old and new weights for training, which can better capture the relationship between the old and new knowledge. And the dynamic growth of the model structure needs to consider the memory and computation consumption problem. Furthermore, we analyze the activation patterns of neurons in the human brain and, in comparison to existing continual knowledge graph embedding models such as LKGE, propose a dual masking mechanism that simulates the guiding role of synaptic excitation and inhibition on neurons.

6.2. Continual Learning Ability

We confirm the continual learning ability of the proposed method by recording the performance of each new time the training is completed in Figure 2. We determine the continual learning capability of the model by recording changes in the MRR metrics as the data grows.

Compared to the other method, our approach leverages an automatic dual-mask filtering mechanism to retain crucial information from old knowledge selectively. When compared to regularization techniques like EWC and SI, our method employs knowledge fusion to enhance the retention of old knowledge based on regularization. Unlike replay-based methods such as GEM and EWR, we utilize a dual-mask mechanism to automatically filter edge information, thereby minimizing the negative impact of irrelevant information on newly acquired knowledge at subsequent time steps. Concurrently, we preserve significant information from old knowledge, mitigating the disastrous forgetting of key historical data. In contrast to the LKGE model, which primarily focuses on the embedding and transfer of local knowledge, our approach fully utilizes graph structural information, enabling a more comprehensive representation of the knowledge.

6.3. Performance and Consumption

We tested the time performance comparison of the compared algorithms in Figure 3. The graph illustrates the performance and consumption of our method on the ENTITY dataset, using the Hits@10 metric. We recorded the training time for each model as a basis for time consumption for model comparison. During new knowledge learning and fusion, the need to aggregate global structural information elevates the model's computational complexity compared to fine-tuned, which predominantly focuses on local information. However, our model's dual-mask mechanism automatically filters out redundant information while retaining critical knowledge. This significantly reduces the model's floating-point calculation complexity. Simultaneously, it eliminates the interference of edge data on new knowledge, enhancing the efficiency of knowledge learning.

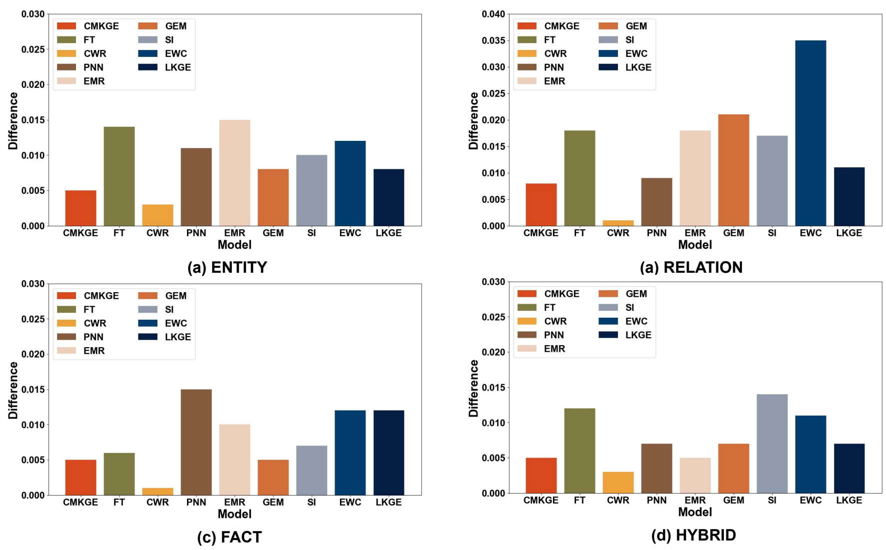

Figure 4: Robustness test on four datasets. "Difference" means the performance degradation of each model after adding wrong data. It is the difference from the original. Although the difference in CWR is smaller, the original performance of the CWR model was much lower than CMKGE, see Table 2.

6.4. Robustness Test

In this subsection, we assess the robustness of the proposed model by introducing a 10% negative triples in the training set. The wrong data is added directly to the training set means that the model does not know what is wrong. The performance change of the CMKGE and the comparison algorithm is documented in Figure 4. Compared with other methods, our method is more robust. When learning new knowledge, we synthesize the representation of the learning knowledge graph by integrating fine-grained local semantic attributes with global structure details. The incorrect triples in the dataset exerts a more significant influence on local attributes. For instance, EWC regularization methods can wrongly penalize

Ablation	ENTITY		RELATION		FACT		HYBRID	
Metrics	MRR	Hit@10	MRR	Hit@10	MRR	Hit@10	MRR	Hit@10
Regularization	0.131(-0.116)	0.269(-0.175)	0.124(-0.086)	0.250(-0.139)	0.103(-0.109)	0.063(-0.327)	0.107(-0.103)	0.147(-0.249)
Global	0.224(-0.023)	0.408(-0.036)	0.184(-0.026)	0.360(-0.029)	0.123(-0.089)	0.278(-0.112)	0.176(-0.034)	0.304(-0.092)
Local	0.107(-0.140)	0.209(-0.235)	0.047(-0.163)	0.107(-0.282)	0.098(-0.114)	0.092(-0.298)	0.034(-0.176)	0.078(-0.318)
Mask	0.227(-0.020)	0.405(-0.039)	0.198(-0.012)	0.375(-0.014)	0.164(-0.048)	0.322(-0.068)	0.190(-0.020)	0.361(-0.035)

Table 3: Ablation experiments.

weight changes in an inappropriate direction. EMR's replay-based approach may retain incorrect information, thereby consistently interfering with the newly learned knowledge. Although the CWR method exhibits good robustness, its initial learning performance is notably low compared to the performance Table 2. On the contrary, while retaining old knowledge, our dual mask considers both a random and weighted perspective. Random masks can mitigate the influence of old knowledge to a degree. Weighted mask employs a learnable threshold to adaptively set mask bounds. CMKGE constrains the impact of erroneous data during new task learning and further masks the retention of incorrect information during knowledge transfer, ultimately enhancing the model's robustness.

6.5. Ablation experiments

Experiments in Table 3 show local learning and regularization to be more useful for methods. a) The local module plays a primary role in knowledge learning, while the global module plays a complementary role. On the one hand, GCN constructs the adjacency matrix unifying all the edge relations, and the edge feature information is not used in the actual training, so it mainly plays an auxiliary role in the global results. On the other hand, in the loss calculation, we set 1 weight for the local learning module and 0.1 for the global learning module to mitigate its influence on learning. b) Regularization plays a primary role in continual learning and the dual masking plays a complementary role. Regularization is directly calculating the loss of distance difference between old and new knowledge to constrain the influence of new knowledge on old knowledge, and plays a major role in old knowledge retention. The dual-masking mechanism mainly uses masks to retain important information in the knowledge and filters irrelevant information to set them to 0, which is further integrated with the new knowledge information. The ratio of masks is low, and its key lies in knowledge filtering and knowledge fusion.

7. Conclusion

Introducing a pioneering framework, this study advances traditional continual learning methods by integrating a dual-mask mechanism, which autonomously identifies and preserves critical knowledge. Besides, harnessing the intrinsic architecture of knowledge graphs, our proposed framework integrates both local attributes and global structural insights to finely tune the delicate balance between plasticity and stability in continual learning scenarios involving knowledge graphs. In contrast to conventional continual learning methodologies and existing knowledge graph continual learning strategies, our proposed framework showcases superior efficacy across all four evaluation benchmarks. Moreover, the model exhibits exceptional resilience and computational efficiency. The perpetual learning capacity of our framework harbors immense promise for domains characterized by ongoing data accumulation, such as recommendation systems and link prediction tasks. In the future, we aspire to transcend traditional methodologies by delving into the intricate architecture of temporal knowledge graphs and integrating them with biologically plausible learning paradigms, such as spiking neural networks, to engineer more efficacious models.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grants No.62204140) and the Key-Area Research and Development Program of Guangdong Province (Grants No. 2021B0909060002).

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

Xiang Chen, Jintian Zhang, Xiaohan Wang, Tongtong Wu, Shumin Deng, Yongheng Wang, Luo Si, Huajun Chen, and Ningyu Zhang. Continual multimodal knowledge graph construction. *arXiv preprint arXiv:2305.08698*, 2023.

Yuanning Cui, Yuxin Wang, Zequn Sun, Wenqiang Liu, Yiqiao Jiang, Kexin Han, and Wei Hu. Lifelong embedding learning and transfer for growing knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4217–4224, 2023.

Angel Daruna, Mehul Gupta, Mohan Sridharan, and Sonia Chernova. Continual learning of knowledge graph embeddings. *IEEE Robotics and Automation Letters*, 6(2):1128–1135, 2021.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Sophie Denève, Alireza Alemi, and Ralph Bourdoukan. The brain as an efficient and robust adaptive learner. *Neuron*, 94(5):969–977, 2017.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

Knud Illeris. A comprehensive understanding of human learning. In *Contemporary theories of learning*, pages 1–14. Routledge, 2018.

Jesper R Ipsen and Andre DH Peterson. Consequences of dale's law on the stability-complexity relationship of random neural networks. *Physical Review E*, 101(5):052412, 2020.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition. *arXiv preprint arXiv:2311.11009*, 2023.

Shuang Liang. Knowledge graph embedding based on graph neural network. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3908–3912, 2023.

Jiajun Liu, Wenjun Ke, Peng Wang, Ziyu Shang, Jinhua Gao, Guozheng Li, Ke Ji, and Yanhe Liu. Towards continual knowledge graph embedding via incremental distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8759–8768, 2024.

Shuwen Liu, Bernardo Cuenca Grau, Ian Horrocks, and Egor V Kostylev. Revisiting inferential benchmarks for knowledge graph completion. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, pages 461–471, 2023.

Haithem Mezni, Djamal Benslimane, and Ladjel Bellatreche. Context-aware service recommendation based on knowledge graph embedding. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5225–5238, 2021.

Bonaventure C Molokwu and Ziad Kobti. Social network analysis using rlvecn: Representation learning via knowledge-graph embeddings and convolutional neural-network. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 5198–5199, 2021.

Dai Quoc Nguyen, Thanh Vu, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A capsule network-based embedding model for knowledge graph completion and search personalization. *arXiv preprint arXiv:1808.04122*, 2018.

Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 3104482–3104584, 2011.

Janna Omeliyanenko, Albin Zehe, Andreas Hotho, and Daniel Schlör. Capskg: Enabling continual knowledge integration in language models for automatic knowledge graph completion. In *International Semantic Web Conference*, pages 618–636, 2023.

Per E Roland. Space-time dynamics of membrane currents evolve to shape excitation, spiking, and inhibition in the cortex at small and large scales. *Neuron*, 94(5):934–942, 2017.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*, 2019.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. Structure-augmented text representation learning for efficient knowledge graph completion. In *Proceedings of the Web Conference 2021*, pages 1737–1748, 2021.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory. *Method and Application*, 2302, 2023a.

Yusong Wang, Dongyuan Li, Kotaro Funakoshi, and Manabu Okumura. Emp: Emotion-guided multi-modal fusion and contrastive learning for personality traits recognition. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 243–252, 2023b.

Yusong Wang, Dongyuan Li, and Jialun Shen. Inter-modality and intra-sample alignment for multi-modal emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8301–8305. IEEE, 2024.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

Fan Yang, Jie Bai, Linjing Li, and Daniel Zeng. A continual learning framework for event prediction with temporal knowledge graphs. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 01–06, 2023.

Peiyan Zhang, Yuchen Yan, Chaozhuo Li, Senzhang Wang, Xing Xie, Guojie Song, and Sunghun Kim. Continual learning on dynamic graphs via parameter isolation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 601–611, 2023.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. *Advances in neural information processing systems*, 32, 2019.

Da Zheng, Xiang Song, Chao Ma, Zeyuan Tan, Zihao Ye, Jin Dong, Hao Xiong, Zheng Zhang, and George Karypis. Dgl-ke: Training knowledge graph embeddings at scale. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 739–748, 2020.

Xiaohan Zou. A survey on application of knowledge graph. In *Journal of Physics: Conference Series*, volume 1487, page 012016. IOP Publishing, 2020.